Bài 5: KHAI THÁC MẪU KẾT HỢP (TT)

I. Mục tiêu:

Sau khi thực hành xong, sinh viên nắm được:

- Thuật toán Vertical Apriori
- Vận dụng thuật toán

II. Tóm tắt lý thuyết:

1. Thuật toán Vertical Apriori:

Thuật toán Aprio áp dụng cho một tập dữ liệu dạng nằm ngang (horizontal) còn thuật toán vertical Apriori áp dụng cho một tập dữ liệu dạng thẳng đứng (vertical).

Thuật toán Vertical Apriori được phát biểu như sau:

```
Algorithm VerticalApriori(Transactions: 𝒯, Minimum Support: minsup)
begin
  k = 1;
  ℱ₁ = { All Frequent 1-itemsets };
  Construct vertical tid lists of each frequent item;
  while ℱₖ is not empty do begin
    Generate 𝒞ₖ₊₁ by joining itemset-pairs in ℱₖ;
    Prune itemsets from 𝒞ₖ₊₁ that violate downward closure;
    Generate tid list of each candidate itemset in 𝒞ₖ₊₁ by intersecting
        tid lists of the itemset-pair in ℱₖ that was used to create it;
    Determine supports of itemsets in 𝒞ₖ₊₁ using lengths of their tid lists;
    ℱₖ₊₁ = Frequent itemsets of 𝒞ₖ₊₁ together with their tid lists;
    k = k + 1;
  end;
  return(∪ᵢ₌₁ᵏ ℱᵢ);
end
```

Ví dụ: cho tập dữ liệu giao dịch được lưu trữ dưới dạng nằm ngang với minsup = 2

TID	Itemset
100	A, B, E
200	B, D
300	B, C
400	A, B, D
500	A, C
600	A, B, C

Ta chuyển dữ liệu thành dạng thẳng đứng

Items	TID list
A	100, 400, 500, 600
B	100, 200, 300, 400, 600
C	300, 500, 600
D	200, 400
E	100

Do minsup = 2 nên tất cả các item xuất hiện chỉ 1 lần trong giao dịch sẽ được loại trừ khỏi tập dữ liệu

Items	TID list
A	100, 400, 500, 600
B	100, 200, 300, 400, 600
C	300, 500, 600
D	200, 400

Tiếp theo ta tạo 1 danh sách chứa 2-item

Item list
A, B
A, C
A, D
B, C
B, D
C, D

Ta kết hợp tất cả item với TID tương ứng

Itemset	TID list
A, B	100, 400, 600
A, C	500, 600
A, D	400
B, C	300, 600
B, D	200, 400
C, D	--

Ta bỏ các 2-item có giá trị support nhỏ hơn minsup = 2:

Itemset	TID list
A, B	100, 400, 600
A, C	500, 600
B, C	300, 600
B, D	200, 400

Ta lặp lại các bước này nhiều lần, tạo 3- item giao dịch và nó có giá trị support nhỏ hơn giá trị của minsup = 2

Items	TID list
A, B, C	600
A, B, D	400
B, C , D	--

nên ta sẽ khởi tạo luật kết hợp dựa vào bước trước đó

Item list
A, B
A, C
A, D
B, C
B, D
C, D

III. Nội dung thực hành:

- Cho CSDL với các giao dịch sau:

TID	Itemset
1	Wine, Chips, Bread, Butter, Milk, Apple
2	Wine, Bread, Butter, Milk
3	Bread, Butter, Milk
4	Chips, Butter, Apple
5	Wine, Chips, Bread, Butter, Milk, Apple
6	Wine, Chips, Milk
7	Wine, Chips, Bread, Butter, Apple
8	Wine, Chips, Milk
9	Wine, Bread, Apple
10	Wine, Bread, Butter, Milk
11	Chips, Bread, Butter, Apple
12	Wine, Butter, Milk, Apple
13	Wine, Chips, Bread, Butter, Milk
14	Wine, Bread, Milk, Apple
15	Wine, Bread, Butter, Milk, Apple
16	Wine, Chips, Bread, Butter, Milk, Apple
17	Chips, Bread, Butter, Milk, Apple
18	Chips, Butter, Milk, Apple
19	Wine, Chips, Bread, Butter, Milk, Apple
20	Wine, Bread, Butter, Milk, Apple
21	Wine, Chips, Bread, Milk, Apple
22	Chips, Bread, Milk

với minsup =3. Cài đặt thuật toán Vertical Apriori và test với tập dữ liệu như trên.